Recommendations for Discouraging, Identifying, and Removing Dirty Data in Survey Research

Justin A. DeSimone

October 4, 2019

THE UNIVERSITY OF ALABAMA®
Culverhouse College of Business
Department of Management

Overview

- Describing dirty data
  - Defining dirty data
  - Distinguishing different types of dirty data
  - Detrimental effects of dirty data

- Dealing with dirty data
  - Determining why respondents provide dirty data
  - Discouraging dirty data
  - Detecting dirty data
  - Deleting dirty data?

What is Dirty Data?

- Construct-irrelevant responses intentionally provided by study participants due to a lack of effort
  - Careless responding
  - Insufficient effort responding (IER)
  - Low-quality data

- Important differentiations:
  - Missing data
  - Outliers
  - Response sets
  - Social desirability

Types of Dirty Data

- Labels:
  - Random responding
  - Straightlining
  - Patterned responding
  - Satisficing
- Three classes of respondents[1,2]:
  - Class 1: Normal responses (89% to 95%)
  - Class 2: “responding in an inconsistent way” (4% to 9%)
  - Class 3: “same response option for many consecutive items” (1% to 2%)

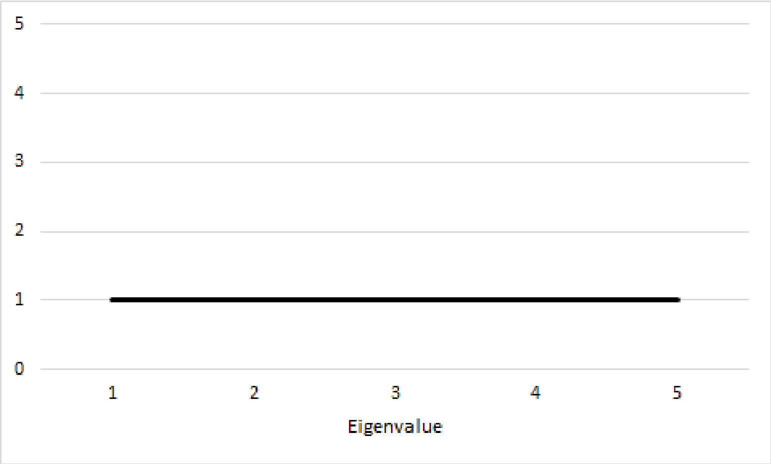
[1]Meade & Craig (2012, quotes on p. 445), [2]Maniaci & Rogge (2014)

Extreme Random Responding

- Random responding occurs when a participant's response to one item has no bearing on his/her response to another item, regardless of the inter-item correlation

r_{ij}	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1	0	0	0	0
Item 2	0	1	0	0	0
Item 3	0	0	1	0	0
Item 4	0	0	0	1	0
Item 5	0	0	0	0	1

- Expected values[3]:
  - Inter-item correlation: 0
  - Alpha: 0
  - First eigenvalue: 1



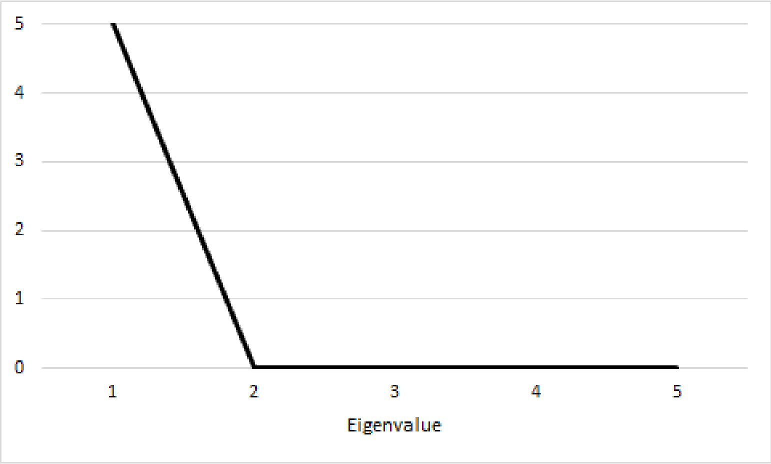
[3]DeSimone et al. (2018)

Extreme Straightlining

- Straightlining occurs when a participant provides identical responses to consecutive items

- Expected values[3]:
  - Inter-item correlation: 1
  - Alpha: 1
  - First eigenvalue: J (#items)

r_{ij}	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1	1	1	1	1	1
Item 2	1	1	1	1	1
Item 3	1	1	1	1	1
Item 4	1	1	1	1	1
Item 5	1	1	1	1	1



[3]DeSimone et al. (2018)

Complications in Real Data

- The expected inter-item correlation for a dataset comprising 100% consistent random responders and straightliners is equal to the proportion of the sample who are straightliners
  - Not all respondents provide dirty data (< 100%)
  - Not all respondents who provide dirty data do so consistently
  - Random responders can also straightline (and vice versa)
  - There may be other forms of dirty data

- Any amount of dirty data is undesirable if it has the potential to influence the results of psychometric or statistical analyses

Influence of Dirty Data on Psychometrics

- Simulated data[3]: Proportion of dirty required to change the conclusions drawn from common psychometric estimates

	α	r_{ij} (SRMR)	λ (CL)	Structure (PCA)
Random responding	65%	30%	15%	85%
Straightlining	n/a	< 5%	10%	20%

- Real data: Effects of screening techniques on psychometric estimates
  - α lower for inattentive respondents[2], though α can increase or decrease depending on which type of dirty data is removed[4]
  - Inter-item and inter-scale correlations change more when removing straightliners than when removing random responders[4]

[2]Maniaci & Rogge (2014), [3]DeSimone et al. (2018), [4]DeSimone & Harms (2018)

Influence of Dirty Data on Study Results

- Dirty data can increase or decrease correlations, depending on the type and distribution of dirty data[2,4]
  - For example, if respondents providing dirty data are more likely to endorse responses near the midpoint of a scale, dirty data may inflate correlations[5,6]
- Counterintuitively, removing participants who provide dirty data can increase statistical power[2]

[2]Maniaci & Rogge (2014), [4]DeSimone & Harms (2018), [5]Credé (2010), [6]Huang et al. (2015)

Why Respondents Provide Dirty Data

- Traditional perspectives[7]:
  - Distraction
  - Lack of interest
  - Laziness
  - Fatigue

- Recent empirical work on the nomological network of dirty data
  - Self-reports
  - Peer-reports, experiments and indirect measures

[7]Krosnick (1991)

Self-report Correlates of Dirty Data

- Five-factor model traits and their facets[6]
  - Neuroticism (positive)
    - Depression, vulnerability
  - Extraversion (negative)
    - Cheerfulness, friendliness
  - Openness (negative)
    - Artistic interests, imagination
  - Agreeableness (negative)
    - Altruism, trust
  - Conscientiousness (negative)
    - Self-efficacy, dutifulness

[6]Huang et al. (2015)

Self-report Correlates of Dirty Data

- HEXACO traits[8]
  - Honesty/Humility (negative)
  - Emotionality (negative)
  - Conscientiousness (negative)

- Dark traits[8]
  - Machiavellianism (positive)
  - Psychopathy (positive)

- Academic outcomes[9]
  - GPA (negative)
  - Class absences (positive)

[8]McKay et al. (2018), [9]Bowling et al. (2016)

Other Correlates of Dirty Data

- Peer-reported personality[9,10]
  - Neuroticism (positive)
  - Extraversion (negative)
  - Agreeableness (negative)
  - Conscientiousness (negative)
- Implicit aggression (positive)[10]
- Questionnaire length (mixed results)[11]

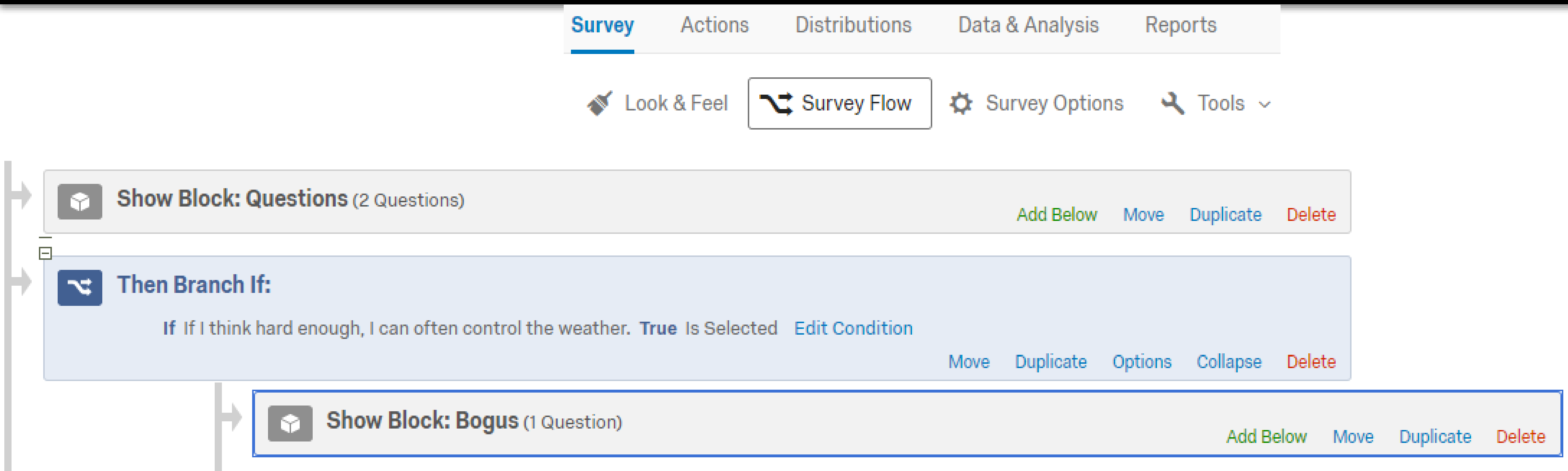
[9]Bowling et al. (2016), [10]DeSimone et al. (in press), [11]Gibson & Bowling (in press)

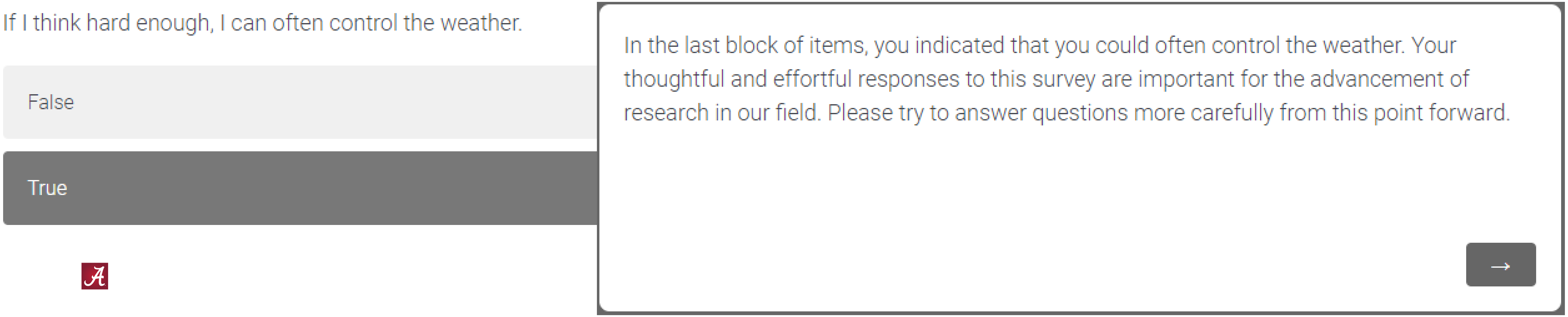
Preventing Dirty Data

- Instructional warnings[1,11]
- Promising rewards[11]
- Asking participants nicely
- Creating a “living survey”

[1]Meade & Craig (2012), [11]Gibson & Bowling (in press)

Living Survey Example





How to Identify Dirty Data

- Data screening indices[12,13]
  - Direct
    - Self-report (or “use me”) questions
    - Bogus or instructed items
  - Archival
    - Response time
    - Longstring/individual response variance (IRV)
    - Semantic synonyms/antonyms
  - Statistical
    - Psychometric synonyms/antonyms
    - Personal reliability
    - Mahalanobis *D*

[12]DeSimone et al. (2015), [13]Curran (2016)

Screening Techniques

Technique	Designed to Capture
Self-report	Confessions of dirty data providers
Bogus or Instructed Items	Random responding and straightlining
Response Time	Random responding and straightlining
Longstring	Straightlining
IRV	Straightlining
Synonyms and Antonyms	Random responding
Personal Reliability	Random responding
Mahalanobis’ *D*	Atypical responding

Data Screening Considerations

- Screening after the results are known (“SARKing”)
- Cutoffs
- Missing data
- Reverse scoring
- Balanced scales
- Screening for multiple types of dirty data
- Maintaining the integrity of scales
- Respondent inconsistency
- Respondent frustration

What to do with Dirty Data?

- Arguments for removing
  - Dirty data is a source of construct-irrelevant variance[4]
  - Dirty data can reduce statistical power[2]
- Arguments for retaining
  - False positives
  - Relationships with focal constructs[10]
- Make an *a priori* choice based on your goal

[2]Maniaci & Rogge (2014), [4]DeSimone & Harms (2018), [10]DeSimone et al. (in press)

Recommendation 1: Anticipate Dirty Data

- Incorporate preventative measures and data screening into study design
- Consider which type(s) of dirty data are the most likely, interesting, or problematic
- Determine screening techniques and cutoffs *before* beginning data collection
- Plan to identify (and possibly remove) dirty data when considering target sample size

Recommendation 2: Screening Techniques

- Use multiple techniques designed to capture multiple types of dirty data
- Always time participant responses, and always consider using other unobtrusive screening techniques
- Use the correct screening techniques for your research design
  - Item wording and direction (balance)
  - Scale length
  - Respondent perception
- Screen conservatively, but scrutinize data

Recommendation 3: Consider Using Dirty Data

- Providing dirty data can be considered an observable behavior
- The tendency to provide dirty data has been demonstrated to correlate with personal characteristics and situational factors
- Dirty data may influence the focal relationships in a study
- Consider modeling or controlling for dirty data as opposed to discarding dirty data

Recommendation 4: Transparency

- Report the details of screening for dirty data
  - Techniques used
  - Cutoffs
  - Deviations from the original plan
  - Determinations of whether or not to remove dirty data
  - Consider sensitivity analysis

- Report results before and after screening
  - How many respondents were flagged for providing dirty data
  - Was dirty data related to any focal constructs
  - How did removing dirty data affect the study results

References

1. Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455.
2. Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83.
3. DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology: An International Review, 67*, 309-338.
4. DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology, 33*, 559-577.
5. Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlation research. Educational and *Psychological Measurement, 70*, 596–612.
6. Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*, 828–845.
7. Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213–236.
8. McKay, A. S., Garcia, D. M., Clapper, J. P., & Shultz, K. S. (2018). The attentive and the careless: Examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Computers in Human Behavior, 84*, 295-303.
9. Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*, 218–229.
10. DeSimone, J. A., Davison, H. K., Schoen, J. L., & Bing, M. N. (in press). Insufficient effort responding as a partial function of implicit aggression. *Organizational Research Methods*.
11. Gibson, A. M., & Bowling, N. A. (in press). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*.
12. DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*, 171–181.
13. Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19.